KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

On the relation between patent citations and patent value

Jurriën Bakker

MSI_1611

On the relation between patent citations and patent value

Towards a more effective model adopting the log-linear form.

Jurriën Bakker
Department of Managerial Economics, Strategy and Innovation
KU Leuven
Naamsestraat 69, 3000, Leuven, Belgium
Jurrien.Bakker@kuleuven.be
Tel: +32 76376201
Fax: +32 16326732

Abstract

This paper reports the results of an analysis of patent citation and patent renewal data, advancing a log-linear relation between patent citations and patent value. A complementary analysis of firms' patent portfolios confirms that modelling the relation between citations and firm value benefits from the adoption of the log-linear form.

Keywords: patent citations, patent value, patent renewal, Tobin's Q

JEL codes: O34

MSC codes : 62N86, 91B84

Acknowledgements

The author wishes to thank Bart Van Looy, Otto Toivanen, Dirk Czarnitzki, and two anonymous reviewers for their insightful comments and invaluable support in completing this paper.

1 Introduction

Forward patent citations are a ubiquitous indicator of the impact and value of patents and, by extension, patent portfolios. Forward citations – i.e. the number of times a patent is deemed relevant prior art by the examiners and/or applicants of later patents – have obtained this position in part due to the extensive validation of this indicator. Validation efforts progressed from early small-scale studies (Carpenter et al. 1981; Trajtenberg, 1991) to larger studies involving large patent sets (e.g. Bessen, 2008; Hall et al. 2005; Gambardella et al. 2008). Moreover, patent citations have been found to correspond to several constructs of value: innovative value (Albert et al. 1991; Arts et al. 2013; Carpenter et al. 1981; Trajtenberg, 1991), private value (Harhoff et al. 1999, 2003; Gambardella et al. 2008) and market value (Belenzon, 2012; Hall et al. 2005).

However, several authors have signaled a disturbing lack of explanatory power when using patent citations to explain patent value (e.g. Gay and Le Bas, 2005; Gittelman, 2008). Furthermore, the distribution of patent citations is skewed and often involves outliers. Consequently, log-linear transformations have been advanced as a solution (e.g. Harhoff et al. 1999; Gambardella et al. 2008). Moreover, a non-linear approach may be required because of the specificities of the patent citation network. For instance, Hung and Wang (2010) found that patent citations follow a rule of preferential attachment. This phenomenon, as first outlined by Barabási and Albert (1991), entails in this case that patents more frequently cite patents that have already been cited regardless of quality considerations. Therefore, later citations could have a lesser value, thus indicating that patent citations do not scale linearly with patent value.

In most studies that involve patent data, patents statistics are grouped. This can be undertaken on the level of a firm's patent portfolio (e.g. Hall et al. 2005) or on national levels (e.g. Neuhäusler and Frietsch, 2012). If patent value does not scale linearly with the sum of patent citations on the individual level, this would also have implications for the group level. Consequently, patent portfolios would have to be calculated differently; simply taking the sum of patent citations to patents in the portfolio would be inadequate.

The continued relevance of patent citations as an important measure of value, as noted in Jaffe and de Rassenfosse (2016), increases the importance of better understanding the relation between patent citations and patent value. This will help in analyses where patent value is modeled as a dependent variable, used as an independent variable, or used as a control. Improvements in using patent citations as a proxy for patent value may also benefit research on patent portfolios and, by extension, modeling the innovative performance of large actors such as firms and countries. Moreover, it secures a better insight into the processes by which patent citations have come to be correlated with patent value.

In this paper, the relevance of a log-linear relationship is demonstrated by relating patent citations to patent renewal data. This is undertaken for patents from both the European Patent Office (EPO) and the United States Patent and Trademark Office (USPTO) because they have different renewal characteristics. Consequently, comparing the analyses of patents from these offices will provide robustness to the results. In a complementary analysis, the derived functional form is assessed by analyzing the value of firm portfolios in an adapted replication of Hall et al. (2005). The results reveal a substantial increase in

explanatory power that may occur when adopting a model specification that reflects the log-linear relationship.

2 Measuring the relation between citations and renewal

2.1 Methods and data

In this analysis, the relationship between patents citations and patent value – as indicated by the decision of patent owners to pay maintenance fees, i.e. patent renewal – is assessed. Patent renewal can be considered an indicator of private patent value (Pakes and Schankerman, 1984; Pakes, 1984; Lanjouw et al. 1998; Harhoff et al. 1999; Thomas, 1999; Hegde and Sampat, 2009; de Rassenfosse and van Pottelsberge de la Potterie, 2013; de Rassenfosse and Jaffe, 2014) since renewal reflects an economic decision on the part of the patent owner. In other words, it registers a minimal private value that the owner assigns to the patent.

Patent data is obtained from the European Patent Office (EPO) PATSTAT fall 2013 database and complemented with renewal data as observed in fee payments to the relevant patent office from its spring 2014 counterpart. The sample was constructed to allow a comparison between EPO and United States Patent and Trademark Office (USPTO) patent applications. Therefore, only DOCDB[1] patent families with granted applications at both the EPO and the USPTO have been included. Using patent applications from 1981 to 2000, the sample includes 538,261 granted EPO applications and 563,603 granted USPTO applications. To better allow a comparison between results obtained for the EPO patents and the USPTO patents, patent citations are observed as citations made to the DOCDB family of the patent by other DOCDB patent families. This measure is comparable across patent offices, unlike the counts of citations to individual patents of different offices, which are affected by different citation practices practiced in different offices and differ considerably (Bakker et al. 2016).

Observing patent renewal for USPTO patents is relatively easy, as one simply has to observe whether maintenance fees have been paid to the USPTO. The USPTO system anticipates three periods of renewal with decisions possible at 4, 8 and 12 years of the patent life. The USPTO renewal time is calculated as the period for which fees have been paid.

Observing patent renewal for EPO patents is more complicated because the EPO has no unitary structure but acts as an intergovernmental organization operating through member state offices. Granted patents are hosted at national offices that subscribe to the EPO (e.g. the Portuguese or the Netherlands patent office). Maintenance payments and renewal decisions are also made at these offices. Thus, an EPO patent may be renewed at one office but abandoned at another. In order to achieve a single EPO renewal indicator, the Single Renewal Approach (SRA) of Van Zeebroeck (2011) is used: the EPO renewal indicator is determined by the longest time a patent has been renewed at any of the national offices subscribing to the EPO convention. Renewal payments at the national offices that

[1] This family groups patents from different offices that have an identical technical content (Albrecht et al. 2010).

subscribe to the EPO are made yearly, and EPO renewal time is therefore calculated as the longest period for which fees have been paid at any of these national offices.

An analysis is performed where a dummy is created for each score level of the DOCDB citation indicator. Here, the dummies are denoted as $DOCDB_i$, where i denotes the citation score. Levels range from 0, 1, 2, 3… 99, 100 … 368+. Patents with a score larger than 100 are grouped in progressively larger clusters consisting of not one, but several, levels of the citation score. For example, patents with a DOCDB citation score of 101, 102 or 103 are grouped in the same cluster. These clusters are treated in the same way as individual citation levels, where their citation score is determined by the central value of the patent citation score of the levels grouped in each cluster. This procedure is undertaken because the density of patents per citation level is otherwise too low for meaningful estimation of the coefficients associated with each level. The number of citation levels per cluster is denoted in Table 1. Finally, all patents with a score of 368 and above (i.e. 9 standard deviation outliers) are grouped together in one category. Because there are few uncited patents, the reference category combines the set of uncited patents with the set where patents are cited only once.

DOCDB citation score	Citation levels per cluster
2-100	1
101-142	3
143-178	5
179-227	7
228-290	9
291-367	11
368+	N/A

Table 1: Number of citation levels grouped together in each cluster as a function of the DOCDB citation score .

Including a set of appropriate control variables ($x_{controls}$) and assuming an independent error term ε, the number of years the patent was maintained ($t_{renewal}$) can be expressed as a function $f()$ of the citation levels $DOCDB_i$ and a constant C:

$$t_{renewal} = f(C + \sum_i \beta_i \cdot DOCDB_i + \sum_{controls} \beta_{controls} \cdot x_{controls}) + \varepsilon$$

In this model, assuming the function $f()$ is correctly chosen to represent the relation between patent value and renewal time, the size of the coefficients β_i should relate to i following the functional form with which patent citations relate to patent value.

Pakes (1986) highlighted a real option approach to the estimation of renewal time by considering that patent renewal not only extends patent protection for a limited time but also provides the option of future extensions. Maurseth (2005) modeled this as a survival problem using a Cox model. This approach rests on the idea that an expected revenue stream can be attributed to a patent in each given year. Whenever the costs of maintaining the patent (are expected to) exceed the revenue stream, the owner of the patent will decide not to continue paying maintenance fees. Because the (modeled) revenue and the costs of maintaining a patent are not constant, the relation between patent value and observed patent life is not linear. Therefore, a Cox survival analysis should better model patent value through patent renewal than a linear regression model such as OLS. The survival model can also take into account the censoring that stems from either the data that is absent due to missing renewal information at the end of the dataset or from the maximum patent lifetime of 20 years.

Cox survival regressions are dependent on the number of distinct possible survival times that can be observed. Multiple objects with the same survival time need to be taken into consideration and a method needs to be employed to resolve these ties. This is especially important for USPTO cases where only three renewal decisions are taken for each patent, resulting in many patents with the exact same survival time. In analyses with many ties, the standard method of resolving them (Breslow, 1974) could yield biased coefficients while the Efron(1974) method has been advocated as an unbiased method (Hertz-Picciotto and Rockhill, 1997; Hsieh, 1995). Therefore, the analyses have been computed using both methods. Small differences were found, but these were not significant in estimating the log-linear fits presented in the results section. In the main analyses, the results of Efron's method are reported since they appear to be less biased than those of Breslow's method (Hertz-Picciotto and Rockhill, 1997).

This paper attempts to construct a framework that applies to all patents. The analysis, therefore, includes control variables concerning the year and the technological class (IPC3 level) of the application, because of the likelihood that these variables affect patent citations as well as renewal probability. Furthermore, it is likely that different applicants have different renewal considerations and write different patents, resulting in different citation characteristics. Thus, the analysis also includes controls that reflect basic characteristics (i.e. type, experience, size and country[2]) of the applicant. This information has been obtained from the harmonized table provided for the EPO PATSTAT database (Magerman et al. 2006; Peeters et al. 2010). Table 2 provides an overview of the definitions and descriptive statistics of the variables used in the analysis.

Name	Description	Mean	Standard deviation	Min	Max
USPTO renewal time	Maximum year for which maintenance fees are paid at the USPTO.	14.59	5.76	4	20
EPO renewal time	Maximum year for which maintenance fees are paid at any national office subscribing to the EPO.	13.12	4.81	2	20
DOCDB$_i$[a]	Dummy indicating whether the DOCDB patent family of the patent is cited i times by other DOCDB families.	22.16	39.02	0.00	3146
Application year[a]	Dummy for the application year of the patent.	1992.59	5.38	1981	2000
IPC3[b]	Dummy variable to indicate if the IPC3 class (e.g. A01) is present in the patent application.	N/A	N/A	N/A	N/A
Applicant experience[c]	Years between filing of current patent and that of the first application filed by the applicant.	36.23	31.12	0	146
Ln(Applt. size)[c]	Logarithm of the total number of patents filed by the applicant.	7.40	3.32	0	12.99
Co-patented	Dummy indicating if the patent has more than 1 applicant.	0.06	0.24	0	1
Applicant type[b]	Type of applicant: company, government, hospital, individual, university or unknown.	N/A	N/A	N/A	N/A
Applicant country[b]	Dummy for the country in which the applicant resided at time of filing the patent.	N/A	N/A	N/A	N/A

Table 2: descriptions and descriptive statistics of USPTO patents in the Cox survival analyses, statistics for EPO patents deviate slightly and are given for EPO renewal. [a] In the case of dummy variables relating to levels of a discrete variable, statistics are given for this variable. [b] indicates partial counts when applicable. Finally, when an application is co-patented, variables with [c] default to the largest and oldest applicants.

2.2 Results

The coefficient estimates (β_i) for the $DOCDB_i$ dummies from the Cox survival regression for the USPTO renewal data are shown in Figure 1. Note that Cox survival regressions estimate hazard (i.e. abandoning

[2] Applicants may come from countries with few patents, which would disrupt the analysis because the dummy variable relating to that country cannot be estimated (well). Therefore, applicant countries with less than 50 patents in the analyses have been grouped together in a separate category. This has affected 819 patents in total.

patents); hence, negative coefficients indicate a higher chance of renewal. The coefficients of the analysis are monotonically decreasing, even though, at higher citation levels, there is greater variance in this relation.

A log-linear function of the form $\beta_i = a + b\ln(c + i)$ is estimated with β_i the size of the dummy coefficient and i the citation score. The log-linear relation depicted fits the relation between the citation scores and the coefficients well (R^2=0.86). As a comparison, a linear curve of the form $\beta_i = a + bi$ has also been estimated, which produces a worse fit (R^2=0.54). Finally, it can be argued that the log-linear fit has one more parameter and would, therefore, have an advantage over the linear estimation. Consequently, a quadratic curve of the form $\beta_i = a + bi + ci^2$ has been estimated as well, and it is a better fit than the linear specification (R^2=0.76). Nevertheless, the log-linear specification is a superior fit to this specification, indicating that the data fit better with a log-linear form than a polynomial with the same number of parameters.

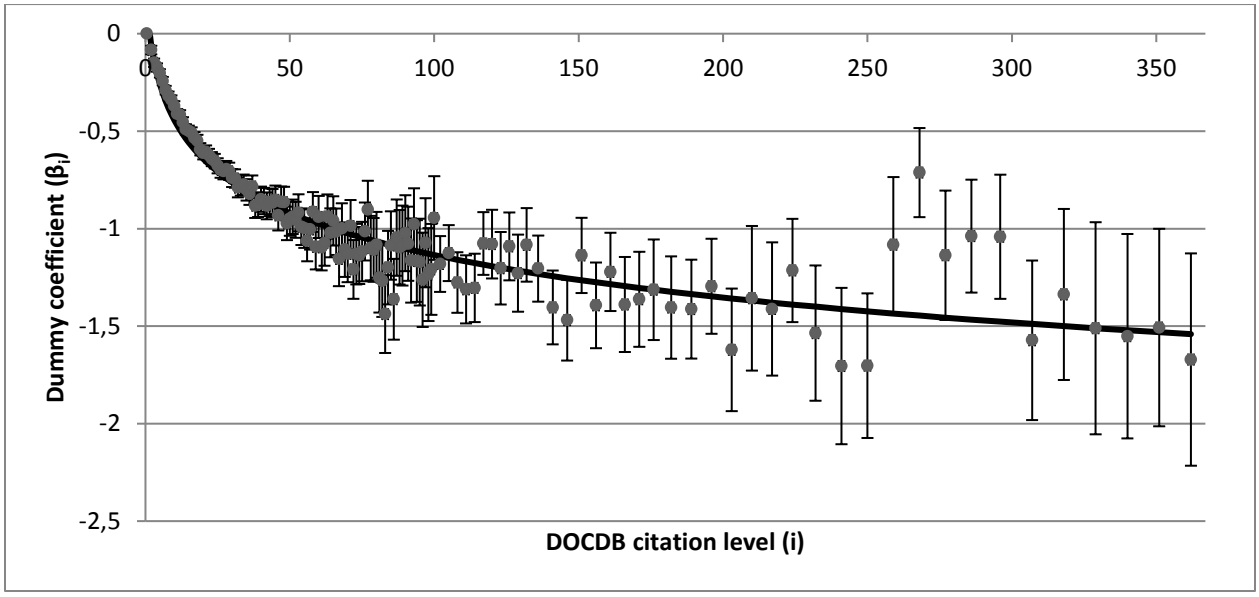

Figure 1: Estimates of the dummy coefficients related to different DOCDB citation scores that were obtained from a Cox survival analysis relating different scores on the DOCDB citation indicator to the maintenance time of a patent at the USPTO. A 95% confidence interval is shown as well as a log-linear fit of $\beta_i = 0.33 - 0.32\ln(1.43 + i)$, which has an R^2 of 0.86.

Similar results are found when repeating the analysis with EPO data; see Figure 2. Here, the fit is even better with R^2=0.93 for the log-linear specification. However, the fits for the other curves also improve with R^2=0.88 for a quadratic curve and R^2=0.78 for a linear curve. Therefore, the found log-linear form appears to be robust with respect to the source of renewal data. Unfortunately, the analysis does not provide a guideline on the optimal offset, given that this parameter varies substantially with a value of 1.43 for the USPTO analysis and a value of 10.80 for the EPO analysis.

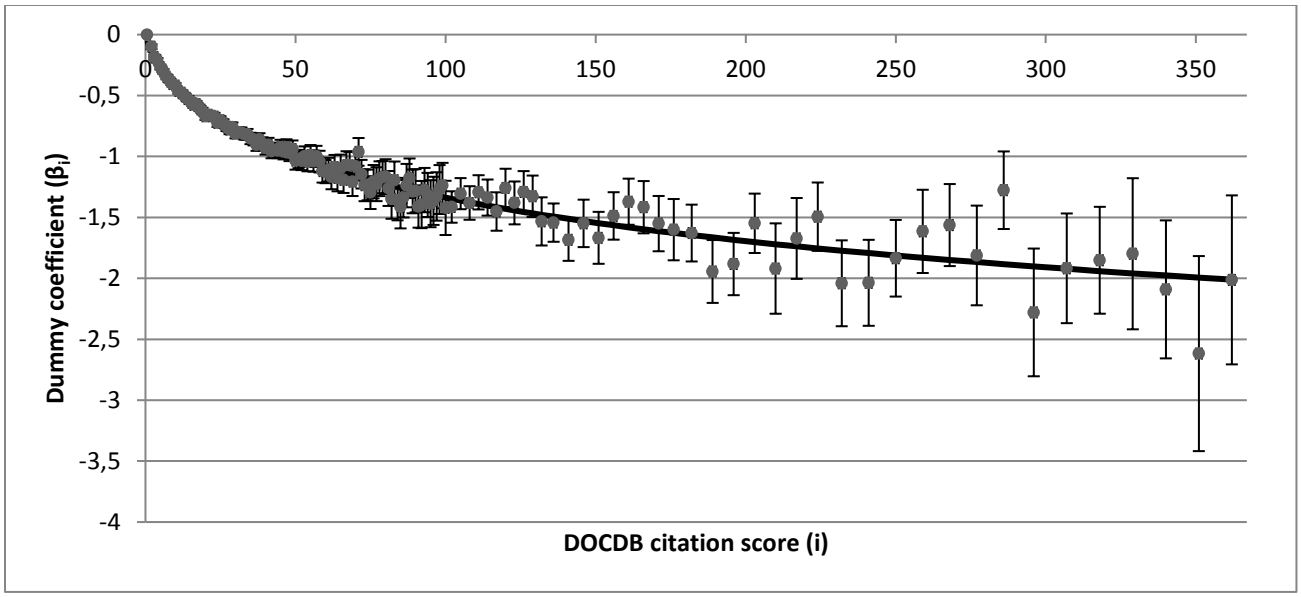

Figure 2: Estimates of the dummy coefficients related to different DOCDB citation scores that were obtained from a Cox survival analysis relating different scores on the DOCDB citation indicator to the maintenance time of a patent at the EPO. A 95% confidence interval is shown as well as a log-linear fit of $\beta_i = 1.27 - 0.55\ln(10.80 + i)$*, which has an* R^2 *of 0.93.*

2.3 Robustness

In this paper, the functional form is estimated using Cox survival analyses. Unfortunately, these analyses rely on the proportional hazards assumption. This assumption states that the hazard function only differs by a constant non-time dependent value between the observed categories. This assumption can be verified by using Schoenfeld (1982) residuals and is violated severely (at $p<0.0001$) for both the USPTO and the EPO analyses.

In consequence, robustness tests using other model specifications have been performed. First, the binary approach from Hegde and Sampat (2009) is adopted, which determines the odds that a patent is renewed by a certain time. This approach has the benefit that it relies on few assumptions concerning the value function of the patent over time. Unfortunately, it also exploits less of the information contained in the renewal data by only observing the patent renewal time at one single time period. In this paper, a binary test is employed where the chance that a patent was renewed until it reached its maximum lifespan (20 years) is estimated using an adaption of equation (1) as shown below.

$$P(t_{renewal} = 20) = P\left(\varepsilon > C - \sum_i \beta_i \cdot DOCDB_i - \sum_{controls} \beta_{controls} \cdot x_{controls}\right)$$

If ε follows a logistic distribution, the β_i coefficients can be estimated using a logistic regression. Because this test relies on patents reaching their maximum lifetime, the sample is confined to patents that have the possibility of reaching it. At the USPTO, where full renewal is decided at 12 years, this includes patents with at least 13 years in the renewal data of spring 2014, which applies to all patents in the original sample. For EPO patents, the renewal decisions are taken yearly; hence, only patents with application years up to 1993 are included in the logistic analysis.

A linear regression analysis is also employed. Here, instead of estimating whether a patent is renewed at a certain point in time, its renewal time $t_{renewal}$ is directly estimated. This estimation can be constructed easily from equation (1), when assuming function $f()$ is linear, leading to the equation listed below.

$$t_{renewal} = C + \sum_{i} \beta_i \cdot DOCDB_i + \sum_{controls} \beta_{controls} \cdot x_{controls} + \varepsilon$$

A linear analysis is interesting because it facilitates use of the richness of the data – i.e. not just whether a patent is renewed but also for how many years – while still not having to rely on the proportional hazards assumption as is the case with the Cox survival analysis. Moreover, this analysis allows for direct computation of the size of the effect that patent citations have on the expected lifetime of a patent. Furthermore, both logistic and Cox survival analyses use a link function that relies on an exponential form. This could affect the form by which the citations correlate with patent renewal. Thus, a linear specification would be helpful in showing that the results found in Section 2.2 are not caused by modeling choices. Unfortunately, this linearization comes with the assumption that the value of a patent and its renewal time are linearly related, which is unlikely, as Section 2.1 explains.

This equation cannot be estimated using OLS because patents can only be renewed up to a certain point (i.e. 20 years). Therefore, the renewal time variable cannot take values greater than 20 years in our data, creating a need to deal with this censoring. Therefore, a Tobit regression analysis is employed, which considers censoring at the maximum lifetime of the patent, i.e. 20 years. Because this analysis has the same selection issues as logistic analysis, the sample is restricted in the same manner.

Finally, a lower bound of the value of a patent can be directly estimated using an interval regression analysis where the intervals are determined by the cumulative renewal fees paid by the owner of the patent. Here again, censoring needs to be considered for patents that reach their maximum lifetime as well as censoring due to limited renewal data. Therefore, the sample restrictions for the binary analysis are also employed here.

The same analyses from Section 2.2 are performed using both the Logit/Tobit and interval regression analyses rather than the Cox survival regression. The same curves relating β_i to i are also estimated. The results from these estimations are presented in Table 3. From the evidence of these results, it is clear that the logarithmic functional form fits best the relation between the estimated coefficients of the dummy β_i and the DOCDB citation score i. The analyses provide fit characteristics that are quite similar, indicating that the relation between patent citations and private value, as indicated by patent renewal, follows the same functional form regardless of the analysis. Therefore, it can be concluded that, of those studied, the log-linear form offers the best description of the functional form by which patent citations relate to patent value.

	USPTO				EPO			
	Cox	**Logit**	**Tobit**	**Interval**	**Cox**	**Logit**	**Tobit**	**Interval**
Linear	0.54	0.57	0.57	0.53	0.78	0.76	0.81	0.80
Quadratic	0.76	0.80	0.80	0.75	0.88	0.88	0.90	0.90
Log-linear	0.86	0.91	0.91	0.87	0.93	0.94	0.94	0.94

Table 3: R^2 of the different fits that relate β_i to i for each analysis at each patent office. It should be noted that the Logit, Tobit and interval analyses for EPO were performed on a smaller sample and are thus not fully comparable with the Cox survival regressions.

2.4 The relation between patent value and patent citations

The results obtained from the main analysis in Section 2.2 and from the robustness analyses are informative in establishing the functional form that relates patent citations to patent value. However, the fits themselves may, in addition, explain the relevance of patent citations in more economic terms. Hence, the fits are presented with an interpretation of their estimated effect in Table 4.

The estimated effect of each additional patent citation is harder to estimate using the log-linear form. Therefore, the effects are given for patents that have double the number of patent citations than a patent with similar scores on the control variables. Given a log-linear fit of $\beta_i = a + b\ln(c + i)$, this translates as $b\ln(2)$. It should be noted that having double the citations should be interpreted using the offset, i.e. the c parameter in the log-linear fit. Therefore ,'doubling' the citations for an uncited EPO patent means adding 14 citations in the case of the Tobit regression.

Office	**Analysis**	**Log-linear fit relating**	**Estimated comparative effect of having double the number of citations than a comparable patent**
USPTO	Cox	$\beta_i = 0.33 - 0.32\ln(1.43 + i)$	Decreased abandonment hazard of 0.22
	Logit	$\beta_i = -0.43 + 0.48\ln(1.12 + i)$	Increased odds of full renewal of 0.33
	Tobit	$\beta_i = -3.15 + 2.86\ln(1.44 + i)$	Increased renewal time of 0.99 years
	Interval	$\beta_i = -1366 + 1642\ln(1.00 + i)$	Increased value of $1137.95
EPO	Cox	$\beta_i = 1.27 - 0.55\ln(10.80 + i)$	Decreased abandonment hazard of 0.38
	Logit	$\beta_i = -1.49 + 0.80\ln(6.49 + i)$	Increased full renewal odds of 0.56
	Tobit	$\beta_i = -8.64 + 3.32\ln(14.05 + i)$	Increased renewal time by 2.30 years
	Interval	$\beta_i = -12738 + 5021\ln(13.67 + i)$	Increased value of €3480.62

Table 4: Fits and economic interpretation of the analyses relating patent citations to patent value

The results listed in Table 4 show that the estimated effect size of having been cited more than a comparable patent is substantial. Estimates show that patent citations confer value that can be measured in years of additional patent life and a value increase of thousands of euros/dollars. It should be noted that patent renewal analyses intrinsically estimate minimum values of patent value. Therefore, the value added by doubling patent citations may very well be much higher. Interestingly, with regard to USPTO patents, this value appears lower than for EPO patents, both in renewal time and patent value. However, the latter is in part due to the lower maintenance fees at the USPTO.

3 Applying the functional form to an econometric analysis

3.1 Introduction

The previous section established a log-linear relation between patent citations and patent value. Sets of patents, i.e. patent portfolios, can also be evaluated using patent citations. The value of a patent portfolio is generally estimated by counting the number of times any patent in the portfolio has been referenced. This practice could be justified on the assumption that the value of a patent portfolio is equal to the sum of the value of its members. The logical conclusion is that, when the value of individual patents is calculated differently, this should have repercussions for the estimation of the value of patent portfolios. Therefore, in this paper, a new method that relies on the found log-linear relation is introduced.

In this log-linear method, the value of a patent portfolio is derived by first computing a log-transformed value for each patent and then computing the sum of these log-transformed values. Because this method better models the relation between patent citations and patent value, it could prove superior to the normal linear method of estimating the value of a patent portfolio, i.e. simply computing the sum of individual patent citations.

A superior method of calculating the value of the patent portfolio would enhance understanding of firm innovative performance, an often-used metric in innovation research. To evaluate the log-linear method, a patent portfolio analysis is presented using both the traditional way of computing the value of a patent portfolio and the proposed log-linear method. For this endeavor, an adapted analysis of Hall et al. (2005) is presented, which relates Tobin's Q, (Q) to stocks of R&D, patents and patent citations.

3.2 Methods and Data

In this chapter, the analysis adapted from Hall et al. (2005) is used to assess which method of evaluating patent portfolios better explains firm performance: the common linear method or the log-linear method that models a log-linear relation between patent citations and patent value. The analysis of Hall et al. (2005) models Tobin's Q as a function of the relative knowledge stock of the firm. This stock is then approximated using the ratio between the R&D stock and the assets of the firm as well as other ratios involving the R&D stock, the patent stock, and the patent citation stock. Controls for year as well as the firm will also be included in the analysis. Therefore, the following equation is estimated:

$$\ln Q_{it} = C_i + C_t + \ln\left(1 + \beta_1 \frac{R\&D_{it}}{A_{it}} + \beta_2 \frac{PAT_{it}}{R\&D_{it}} + \beta_3 \frac{CITES_{it}}{PAT_{it}} + D_{R\&D_{it}=0}\right) + \varepsilon_{it}$$

Here C_i and C_t denote constants of firm i and time t while A_{it} denotes the total assets. $R\&D_{it}$, PAT_{it} and $CITES_{it}$ denote respectively the R&D, patent and citations depreciated stock. $D_{R\&D_{it}=0}$ denotes a dummy for firms with no reported R&D expenditures at time t. When this dummy is equal to 1, the ratio $\frac{PAT_{it}}{R\&D_{it}}$ is set to 0 if the R&D stock, $R\&D_{it}$, is equal to 0. There are also cases for which the patent stock is equal to 0; these are not used in the analysis. Finally, ε_{it} represents a random error.

The non-linear analysis that follows from equation 4 is presented along with a linearized version, which assumes $\ln(1 + x) \approx x$. This linear analysis has the benefit of facilitating a fixed effects approach, which

is not possible with the non-linear analysis, as noted in Hall et al. (2005). In the non-linear analyses, C_i is approximated using sector dummies of the Standard Industry Classification (SIC).

For this analysis, USPTO applications (since the sample mainly concerns US firms) combined with DOCDB citation information are matched to a random sample of patenting firms with at least 100 patents listed in PATSTAT, and that are listed in the Compustat database. For the resulting sample of 1092 firms, financial data is considered from the years between 1981 and 2005. In this paper, citation stock increases are modeled using a linear model as well as the log-linear model previously specified.

The sample was constructed as follows: only firms that have a continuous presence in at least two periods in the dataset were used. Moreover, in order to accurately compare patenting firms, only observations of firms that have a non-zero patent stock – i.e. observations of firms that have at least one patent in the current or any previous period – are used.[3] Finally, Tobin's Q was not known for all observations in the resulting dataset, leading to the removal of 1890 observations.

The R&D stock was initialized as the R&D expenditure for the first year in which a firm enters the sample divided by 0.23, in a procedure similar to Hall (1990) and Hall et al. (2005). The patent stock and the citation stocks are not initialized because the full patenting activity of all firms is observed for 30 years prior to the first year of the sample using the EPO PATSTAT database. Finally, all stocks are depreciated by 15% each year, in line with Hall et al. (2005). The descriptive statistics of the sample are detailed in Table 5.

Variable	Description	Number of Observations	Mean	Standard Deviation	Min	Max
Ln (Tobin's Q)	Natural log of market value divided by total assets	13044	0.39	0.78	-6.23	4.63
Year	Book year of the firm, application year of the patents	13044	1995	7.39	1980	2005
R&D stock	The current stock of R&D expenses ($M)	13044	631	2091	0.00	29814
Total Assets	The total assets of the firm ($M)	13044	4841	24430	0.04	658800
D(R&D=0)	Dummy to indicate no R&D expenses in that year	13044	0.116	0.32	0	1
Patent stock	The current stock of USPTO patents	13044	239	843	0.00	14649
Citation stock $\sum \boldsymbol{CIT}$	The current citation stock, calculated using the linear method	13044	4402	15156	0.02	226776
Citation stock $\sum \mathbf{ln}(\mathbf{1}+\boldsymbol{CIT})$	The current citation stock, calculated using the log-linear method	13044	579	2038	0.01	31932

Table 5: Descriptive statistics of the variables used in the horse race regressions. Below each citation stock is listed the formula used to create it, where CIT refers to the citation score of an individual patent. R&D stock and total assets are adjusted for inflation using USBLS(2016) data (1983=100). All stocks are calculated with a 15% depreciation rate.

3.3 Results

The results of the fixed effects linear models are shown in Table 6. The increase in R^2 shows that citation indicators using the log-linear transformation perform better – with an increase of 3.4% in explained variance – at explaining log Tobin's Q than the citation indicators without the use of the logarithmic transformation. This represents a substantial increase of 70% in added explained variance by introducing

[3] A comparative analysis that included observations for firms with no patent stock as well as a dummy controlling for this occurrence yielded very similar results to the analyses presented in this paper.

a citation indicator to explain company performance. Therefore, this analysis shows the potential of applying the log transformation to portfolio analysis, while simultaneously providing external validity to the findings in Chapter 2.

The non-linear analysis of Hall et al. (2005) was also performed: see analyses 4, 5 and 6. Applying their analysis to this paper's sample produces very similar results, with one exception: the ratio of patent stock over R&D stock, representing patenting efficiency, is negative. The likely cause is the inclusion of several firms for which R&D expenditure is not listed in the COMPUSTAT database and for which this ratio is recorded as 0. Analyses excluding these firms produce a positive coefficient for this ratio. The linear and the log-linear specifications of the citation stock perform very similarly in the non-linear analysis: there is only a difference of 0.007 in their R^2. Using the log-linear form only adds 2.9% in added explained variance.

Including firm dummies instead of SIC dummies give very similar results, as can be seen from analyses 7,8 and 9. However, here the linear specification performs slightly (0.006) better. This represents a decrease of 17% in explained variance when using the log-linear method as opposed to the classical method. This analysis, therefore, demonstrates that that the log-linear method of citation counting does not always deliver improvement, but it produces adequate results nonetheless.

	(1) Fixed effects	(2) Fixed effects	(3) Fixed effects	(4) Non-linear	(5) Non-linear	(6) Non-linear	(7) Non-linear	(8) Non-linear	(9) Non-linear
R&D[a]/ Total assets	0.0154**	0.0156**	0.0156**	0.100**	0.145**	0.237*	0.0370	0.0439	0.0558
	(0.00553)	(0.00550)	(0.00546)	(0.0343)	(0.0463)	(0.0922)	(0.0271)	(0.0316)	(0.0406)
Patents[a]/R&D[a]	-0.000883**	-0.000743	-0.000781*	-0.000456***	-0.000545**	-0.00114***	-0.000509***	-0.000544**	-0.000723***
	(0.000297)	(0.000387)	(0.000324)	(0.000110)	(0.000181)	(0.000222)	(0.000133)	(0.000192)	(0.000180)
D(R&D$_{it}$=0)	0.0880	0.0811	0.0767	0.0122	0.0422	0.0620	0.0893	0.0960	0.127
	(0.0668)	(0.0666)	(0.0652)	(0.0519)	(0.0588)	(0.119)	(0.0775)	(0.0825)	(0.105)
Citations[a]/Patents[a] $\sum \mathbf{CIT}$		0.00213***			0.0119***			0.00457**	
		(0.000579)			(0.00181)			(0.00155)	
Citations[a]/Patents[a] $\sum \ln(\mathbf{1}+\mathbf{CIT})$			0.124***			0.630***			0.159*
			(0.0308)			(0.152)			(0.0642)
Constant	0.258***	0.204***	-0.0601	0.390***	0.0954	-0.540**	-0.324***	-0.355***	-0.567***
	(0.0302)	(0.0333)	(0.0834)	(0.117)	(0.122)	(0.201)	(0.0153)	(0.0177)	(0.0877)
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm controls	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes
SIC dummies	No	No	No	Yes	Yes	Yes	No	No	No
N	13044	13044	13044	13044	13044	13044	13044	13044	13044
Nr. Firms	1092	1092	1092	1092	1092	1092	1092	1092	1092
Nr. SIC	214	214	214	214	214	214	214	214	214
R^2	0.075	0.125	0.160	0.4666	0.4905	0.4912	0.7028	0.7052	0.7046

Table 6: Horse race regressions explaining Ln(Tobin's Q). Variables with $\boldsymbol{a}$ *represent stocks with a 15% depreciation rate. Below each citation stock is listed the formula used to create it, where CIT refers to the citation score of an individual patent. Cluster-robust standard errors are reported in parentheses and asterisks indicate statistical significance with: * p < 0.05, ** p < 0.01, *** p < 0.001.*

4 Conclusion

The main result of this paper is that patent citations display a log-linear relation with patent value. Therefore, researchers are advised to take this relation into consideration when using patent citations to approximate patent value. The fits obtained from the renewal analysis show that patents with double the number of citations than comparable patents have an increased value of $1137.95 in the case of USPTO patents and €3480.62 in the case of EPO patents. The results of the firm analysis indicate that, at least in some economic models, it may be better to first apply a log transformation to the citation count of an individual patent before computing the sum. Doing so may yield an improvement of up to 70% in added explained variance. Yet, in another analysis, the classical way of calculating patent citations has proved slightly superior. For that reason, the log-linear transformation should be used with caution.

When using a logarithmic functional form, the explanatory power of the citation indicator improves. Yet, much unexplained variance remains. Therefore, we should continue to keep in mind the limited ability of patent citations to approximate patent value. Moreover, the found functional form reflects the relation between patent citations and private value, but it may not hold true for other value constructs such as the social value and (knowledge) impact of a patent. Hence, researchers should be careful when applying the findings of this paper to approximate other constructs of patent value.

5 References

Albert, M. B., Avery, D., Narin, F., & McAllister, P. (1991). Direct validation of citation counts as indicators of industrially important patents. *Research Policy*, 20(3), 251-259.

Albrecht, M. A., Bosma, R., van Dinter, T., Ernst, J. L., van Ginkel, K., & Versloot-Spoelstra, F. (2010). Quality assurance in the EPO patent information resource. *World Patent Information*, 32(4), 279-286.

Arts, S., Appio, F. P., & Van Looy, B. (2013). Inventions shaping technological trajectories: do existing patent indicators provide a comprehensive picture?.*Scientometrics*, 97(2), 397-419.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509-512.

Bakker, J., Verhoeven, D., Zhang, L., & Van Looy, B. (2016). Patent citation indicators: One size fits all? *Scientometrics*, 106(1), 187-211.

Belenzon, S. (2012). Cumulative Innovation and Market Value: Evidence from Patent Citations. *The Economic Journal*, 122(559), 265-285.

Bessen, J. (2008). The value of US patents by owner and patent characteristics. *Research Policy*, 37(5), 932-945.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 89-99.

Carpenter, M. P., Narin, F., & Woolf, P. (1981). Citation rates to technologically important patents. *World Patent Information*, 3(4), 160-163.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American statistical Association*, *72*(359), 557-565.

Gambardella, A., Harhoff, D., & Verspagen, B. (2008). The value of European patents. *European Management Review*, 5(2), 69-84.

Gay, C., & Le Bas, C. (2005). Uses without too many abuses of patent citations or the simple economics of patent citations as a measure of value and flows of knowledge. *Economics of Innovation and New Technology*, 14(5), 333-338.

Gittelman, M. (2008). A note on the value of patents as indicators of innovation: *Implications for management research*. The Academy of Management Perspectives, 22(3), 21-27.

Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of Economics*, 16-38.

Harhoff, D., Narin, F., Scherer, F. M., & Vopel, K. (1999). Citation frequency and the value of patented inventions. *Review of Economics and Statistics*, 81(3), 511-515.

Hegde, D., & Sampat, B. (2009). Examiner citations, applicant citations, and the private value of patents. *Economics Letters*, 105(3), 287-289.

Hertz-Picciotto, I., & Rockhill, B. (1997). Validity and efficiency of approximation methods for tied survival times in Cox regression. *Biometrics*, 1151-1156.

Hsieh, F. Y. (1995). A cautionary note on the analysis of extreme data with Cox regression. *The American Statistician*, 49(2), 226-228.

Hung, S. W., & Wang, A. P. (2010). Examining the small world phenomenon in the patent citation network: a case study of the radio frequency identification (RFID) network. *Scientometrics*, 82(1), 121-134.

Jaffe, A. B., & De Rassenfosse, G. (2016). Patent citation data in social science research: Overview and best practices (No. w21868). *National Bureau of Economic Research*.

Maurseth, P. B. (2005). Lovely but dangerous: *The impact of patent citations on patent renewal. Economics of Innovation and New Technology*, 14(5), 351-374.

Magerman, T., Van Looy, B., & Song, X. (2006). Data production methods for harmonized patent indicators: Patentee Name Harmonization. *EUROSTAT Working Paper and Studies*, Luxembourg.

Pakes, A. (1986). Patents as options: Some estimates of the value of holding European patent stocks (No. w1340). *National Bureau of Economic Research*.

Peeters, B., Song, X., Callaert, J., Grouwels, J., & Van Looy, B. (2010). Harmonizing harmonized patentee names: an exploratory assessment of top patentees. *Eurostat Working Paper*.

De la Potterie, B. V. P., & Van Zeebroeck, N. (2008). A brief history of space and time: The scope-year index as a patent value indicator based on families and renewals. *Scientometrics*, 75(2), 319-338.

De Rassenfosse, G., & Van Pottelsberghe de la Potterie, B. (2013). The role of fees in patent systems: Theory and evidence. *Journal of Economic Surveys*, 27(4), 696-716.

De Rassenfosse, G., & Jaffe, A. B. (2014). Are patent fees effective at weeding out low-quality patents? (No. w20785). *National Bureau of Economic Research*.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1), 239-241.

Thomas, P. (1999). The effect of technological impact upon patent renewal decisions. *Technology Analysis & Strategic Management*, 11(2), 181-197.

United States Bureau of Labor Statistics(2016). Consumer price index- All Urban Consumers, not seasonally adjusted, series id: CUUR0000SA0. http://data.bls.gov/timeseries/CUUR0000SA0. Accessed 1 September 2016.

Van Zeebroeck, N. (2011). The puzzle of patent value indicators. *Economics of Innovation and New Technology*, 20(1), 33-62.

FACULTY OF ECONOMICS AND BUSINESS
DEPARTMENT OF MANAGERIAL ECONOMICS, STRATEGY AND INNOVATION
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 67 00
fax + 32 16 32 67 32
info@econ.kuleuven.be
www.econ.kuleuven.be/MSI